Journal of Educational Research & Policy Studies
Spring 2009, Vol. 9, No. 1, pp. 81-100

NORMES

When Measures Change over Time: A Workable Solution for Analysing Change and Growth across Multiple Waves

Jennifer E. V. Lloyd
University of British Columbia

Bruno D. Zumbo
University of British Columbia

Linda S. Siegel
University of British Columbia

In the past 20 years, the analysis of individual change has become a key focus in educational research. There are several parametric analyses that centre upon quantifying change. Some researchers state that such analyses should only occur if the measure itself remains completely unchanged across waves, arguing that it is not possible to link or connect the scores, either methodologically or conceptually, of measures whose content, wording, response categories, or response formats vary across waves. Because it is not always possible or warranted to use the exact same measure over time, however, it is vital to explore more fully the problem of analysing change and growth with measures that vary across waves. To this end, the primary objective of this paper is to expand upon the statistical work of Lloyd and Zumbo (2007) by introducing the non-parametric hierarchical linear model (NPAR-HLM), a workable solution to the problem of analysing change/growth with measures that change over multiple waves. An example of the implementation of the solution is provided, as is a discussion of the solution's assumptions, strengths, and limitations.

Introduction

Individual change is the subject of significant attention in education. The analysis of such change is aimed at quantifying the amount by which individuals grow, mature, improve, and progress over time. By measuring and tracking changes over hours,

days, weeks, months, and even years, it is possible to reveal the temporal nature of development (Singer & Willett, 2003).

Repeated measures analyses – a class of parametric methodologies that centre upon quantifying change over time – are often characterised by one set of individuals being measured more than once on the same (or commensurable) dependent variable. Some researchers state that such analyses should only occur if the measure itself remains *completely unchanged* across waves, arguing that it is not possible to link or connect the scores, either methodologically or conceptually, of measures whose content, wording, response categories, or response formats vary across waves (von Davier, Holland, & Thayer, 2004; Willett, Singer, & Martin, 1998). Because it is not always possible or warranted to use the exact same measure over time, however, it is vital to explore more fully the problem of analysing change and growth with measures that vary across waves. In fact, the practise of analysing change over time using measures that themselves change across waves is commonly found in educational research.

Lloyd and Zumbo (2007) introduce a solution to the problem of analysing change with *time-variable measures* – that is, measures whose content, wording, response categories, or response formats vary in some systematic way across waves. Their solution, called the *non-parametric difference score (NPAR-Diff)*, involves using ranks in the place of original scores in regular parametric analyses. It is this use of ranks, instead of original scores, that makes the *NPAR-Diff* a non-parametric solution. Designed to explore change across *two waves* (e.g., pretest-posttest designs), the *NPAR-Diff* was created for applied researchers and can be implemented easily in everyday research settings.

Objective

It is recognised increasingly that there is a need to explore solutions for handling the problem of analysing change and growth when one uses time-variable measures across multiple waves. Therefore, the primary purpose of this paper is to expand upon the statistical and technical work laid out by Lloyd and Zumbo (2007), applying the non-parametric solution to the context of analysing change and growth across *multiple (i.e., three or more) waves*. In the next section, we further set the context of this paper by discussing three research scenarios in which one may utilise repeated measures analyses.

Setting the Context

Using Repeated Measures Analyses: Three Research Scenarios

There are three research scenarios in which repeated measures can be utilised:

Scenario 1: Exact same measure across waves. In this scenario, one's construct of choice makes it possible to use and re-use the *exact same* measure across all

waves, irrespective of the ever-changing age, cognitive development, and personal and scholarly experiences of the test-takers. The measures' content, item wording, response categories, and response formats do not change whatsoever across waves.

Scenario 2: Linkable time-variable measures. Time-variable measures are those whose content, wording, response categories, and/or response formats vary across waves in repeated measures designs. In this scenario, although the time-variable measures are not completely identical across waves, there is at least one anchor item *shared* by each of the measures, on whose scores traditional test linking or equating techniques (e.g., vertical scaling, item response theory techniques) can be performed (Kolen & Brennan, 2004).

Scenario 3: Non-linkable time-variable measures. This scenario involves using measures whose content, item wording, response categories, and/or response formats may vary *completely* across waves (Mislevy, 1992). Take, for example, a mathematics achievement test administered at Grades 4, 7, and then 10: The measure administered at Grade 4 cannot be same as the measure at Grade 7 and, in turn, the Grade 7 measure cannot be the same as the Grade 10 measure. If they were the same, the reliability and validity of the test scores would be compromised, likely rendering the study useless (Singer & Willett, 2003).

Although Scenario 3 has been characterised as the situation in which one's measures share no common items across waves, it is also possible to encounter Scenario 3 in two additional situations: (a) when one has linkable time-variable measures but a small sample size, or (b) when one does not have the ability to compare the sample's scores to those of a norming group. In the case of small sample sizes, it is not necessarily advisable to link or connect the scores of measures, even if the measures share common items.

As Scenario 1 (exact same measure across waves) illustrates, certain constructs can indeed be measured using the exact same measure across all testing occasions – irrespective of the ever-changing age, cognitive development, and personal and scholarly experiences of the test-takers. As Scenario 2 (linkable time-variable measures) and particularly Scenario 3 (non-linkable time-variable measures) depict, however, there are often situations in which one's construct of choice makes using and re-using the exact same measure across waves unreasonable and even impossible. Kolen and Brennan (2004) remind readers that, when test forms cannot be made to be identical, it may be impossible to equate. So what is one to do, then, if the use of time-variable measures is necessary and unavoidable?

To this end, the problem motivating this paper surrounds the analysis of change and growth with non-linkable time-variable measures (i.e., Scenario 3). Of course, Scenario 2 is also relevant because the problem of handling linkable time-variable measures has been relatively under-studied and under-documented in the test linking and change/growth literatures. Many of the strategies currently being used in the change/growth literature as means of handling the motivating problem, such as vertical scaling and item response theory techniques (see Kolen & Brennan, 2004), are often only useful to

large testing organisations that have access to very large numbers of test-takers and/or expansive item pools. Practically speaking, researchers in everyday research settings must often make due with small sample sizes and novel, never-before-used measures. It is for this reason that this paper introduces a *workable* solution to the problem of analysing change/growth with time-variable measures administered over multiple waves – one that can be implemented easily in *everyday* research settings. In the next section, the value of merging non-parametric and parametric approaches is discussed.

Bridging the Non-Parametric/Parametric Gap

Researchers often underestimate or overlook the utility of non-parametric statistical methods. Introductory statistics course instructors frequently isolate 'rank speak' into separate units that appear to students and/or day-to-day researchers to be disconnected from the general flow of the material (Conover & Iman, 1981), and textbook writers frequently convey the erroneous impression that non-parametric analyses are 'always' less powerful than their parametric counterparts (Zimmerman & Zumbo, 2005). There can, however, be great utility in bridging the gap between non-parametric and parametric statistical methods, as noted by Conover and Iman (1981), Conover (1999), and Zimmerman and Zumbo (1993, 2005).

Indeed, this gap has already been bridged in various contexts, such as the Spearman correlation coefficient (Conover, 1999; Conover & Iman, 1981). Seldom, however, has this bridge been extended to the problem of analysing change/growth, particularly with time-variable measures. By expanding upon the previous research of Conover and his colleagues – whereby the bridge between non-parametric and parametric methodologies is extended to the problem of analysing change/growth with time-variable measures – it is now possible to perform a new change/growth analysis: the *non-parametric hierarchical linear model (NPAR-HLM)* for multi-wave data that cannot be linked or equated. Justification for this strategy is further provided by Lloyd (2006) and Lloyd and Zumbo (2007), and by Braun (1988), who describes a percentile-based method by which to bridge the non-parametric/parametric gap, both within proper longitudinal designs and repeated cross-sectional (pseudo-longitudinal) designs.

The Non-Parametric HLM: The Conover Solution for Multi-Wave Data

The *NPAR-HLM* involves rank transforming (or ordering) individuals' original longitudinal test scores *within wave* pre-analysis, and then using these ranks *in the place of* the original scores in subsequent hierarchical linear modelling analyses. We refer to the general approach of converting original scores into ranks pre-analysis as the Conover solution, in recognition of the seminal work of W. J. Conover (e.g., Conover, 1999; Conover & Iman, 1981) whose research not only inspired the *NPAR-HLM*, but also provides evidence for the solution's statistical viability.

A *rank* depicts the position of a test-taker on a variable relative to the positions held by all other test-takers. *Ranking* or *rank transforming* refers to the process of

transforming a test-taker's original score to a rank relative to other test-takers – suggesting a one-to-one function *f* from the set $\{X_1, X_2, \ldots, X_N\}$ (the sample values) to the set $\{1, 2, \ldots, N\}$ (the first *N* positive integers) (Zimmerman & Zumbo, 1993).

For example, if Test-taker *X* earned a score of 12, Test-taker *Y* earned a score of 13, and Test-taker *Z* earned a score of 14, then the test-takers' respective ranks would be 1, 2, and 3 (where a rank of 1 is assigned to the test-taker with the lowest score).[1] As Zimmerman and Zumbo (2005) remind readers, inasmuch as test-takers' scores are represented in terms of their position relative to other test-takers in the same wave, a rank is similar to a percentile score. A *percentile score* is a type of rank that represents an original score as the percentage of test-takers in an external norming group whose score falls below that score. Unfortunately, referring to the scores of a norming sample is not always practical or possible.

In general, researchers can effectively use ranks in situations in which regular statistical assumptions (e.g., normality) are not or cannot be met (see, for example, Beasley & Zumbo, 2003; Conover, 1999; Zimmerman & Zumbo, 1993). Zimmerman and Zumbo (1993, 2005) note that using ranks in parametric analyses can, in fact, often *improve* the validity and power of significance tests for non-normal distributions. Moreover, ranks used in parametric analysis often produce similar results to those of traditional non-parametric tests (Zimmerman & Zumbo, 2005), and can be used in multi-wave designs.

Assumptions of the NPAR-HLM Solution

As with any methodological tool, the *NPAR-HLM* solution comes with its own set of assumptions.[2] First, the scale for the measures' original scores must be ordinal, interval, or ratio. If so, then the original scores can be converted into ranks in a meaningful fashion. If, however, the scale is categorical, then there is no way of meaningfully converting the original scores into ranks.

Second, there must be *heterogeneous change* in the ranks across waves, meaning that all test-takers do *not* change the same amount across waves. Imagine that Test-Taker *X* earns a rank = 1 across all waves and Test-Taker *Y* earns a rank = 2 across all waves. In this example, both test-takers' rank-based change scores equal zero, suggesting *homogeneous* change which, for reasons outlined by Zumbo (1999), is not permissible in change-over-time analyses. It should be noted that homogeneous change is not a problem endemic to only the Conover solution. Homogeneous change also degrades the value of calculating simple difference scores (Zumbo, 1999).

Finally, the *NPAR-HLM* solution requires that a commensurable (or comparable or similar) construct is measured across all waves of the study. Although the means by which one can assess the commensurability of time-variable measures is beyond the scope of this paper, commensurability generally means that the same primary dimension or latent variable is driving the test-takers' responses across waves (Lloyd & Zumbo, 2007). A latent variable is an unobserved variable that accounts for the correlation among one's observed or manifest variables. Ideally, psychometricians

design measures such that the latent variable that drives test-takers' responses is a representation of the construct of interest.

How Utilising the NPAR-HLM Changes the Research Question

It is important to note that, when one applies the *NPAR-HLM* to the problem of analysing change/growth with time-variable measures, one changes the research question being investigated: One is no longer investigating the factors that contribute to test-takers' performance growth across the waves. Instead, one is examining the factors that contribute to the consistency, or lack thereof, of test-takers' rank order across waves. Put another way, the *NPAR-HLM* allows for exploration in changes in relative rank across waves, rather than rates of change.

Change in relative rank across waves may appear, at first blush, to be a less preferred focus than performance growth. When one deals with time-variable measures, however, performance growth results based on original scores are *not interpretable*, essentially because the various measures' scores have not been placed onto any sort of common metric. Therefore, there is "no way of assigning a useful interpretation to observed differences in gains" (Braun, 1988, p. 174). The fact that the rank transformation changes the research question is not new to research methodologists. For example, the Spearman correlation coefficient measures the degree of monotonic relationship between two variables, as compared to the Pearson correlation coefficient which measures the degree of a linear relationship between two variables.

Introducing the NPAR-HLM in the Context of an Example

Imagine that a researcher is interested in exploring whether or not there are gender differences in the *rank-based* longitudinal reading achievement scores of a group of test-takers. Obtained were the original (raw) scores on the Stanford Diagnostic Reading Test (SDRT), a standardised test of reading comprehension designed by Harcourt Assessment (Karlsen & Gardner, 1978-1996).

Data were collected for 653 children (n_{female} = 336, n_{male} = 317), each of whom was assessed across five waves: Grades 2, 3, 4, 5, and 6, inclusive. As discussed more fully in a later section, test-takers missing one or more waves of SDRT data were excluded from analyses.

The SDRT administration involves each test-taker's receiving a booklet, reading the short passages within the booklet, and providing responses to multiple-choice questions based on the reading in a prescribed time limit (Lesaux & Siegel, 2003). Because test-takers' reading comprehension changes with time, the SDRT has been changed developmentally (i.e., across waves) by the test developer (Karlsen & Gardner, 1978-1996). It should be noted that the SDRT is an example of a *non-linkable* time-variable measure, because the actual paragraphs read and the questions asked differ at each grade level.

As Table 1 and Figure 1 depict, the descriptive statistics for each wave of SDRT original scores vary widely.

Because this example involves data collected over three or more waves, it is possible to conduct a two-level HLM analysis of change: Test-takers' rank-transformed test scores (the Level 1 outcome variable) are nested within test-takers (the Level 2 grouping units), with each test-taker's gender serving as the Level 2 *predictor variable*. It should be noted that gender was chosen as the Level 2 predictor for illustrative reasons only.

The key to implementing the *NPAR-HLM* is that one must first rank transform the data *within wave*, with the mean rank being assigned to ties – a process that requires data being entered in the spreadsheet in *person-level* format, in which one row represents one test-taker, with time-related variables represented along the horizontal of the spreadsheet. After the rank transformation has been completed, one can then transpose the data into *person-period* format, in which each test-taker has multiple rows (one for each wave) and, in turn, proceed with the change/growth analysis.

Figure 2 illustrates that Test-Taker *X*, for example, earns a rank of 2 for Wave 1 (Grade 2) because his original Wave 1 score (30) is between those of Test-Taker *Y* (27, Rank = 1) and Test-Taker *Z* (36, Rank = 3). Recall from an earlier section that a rank of 1 is assigned to the test-taker with the lowest within-wave score.

Statistical Models and Equations

When dealing with nested data, two sets of analyses are performed: *unconditional* and *conditional*. By doing so, one can then determine what improvement in the prediction of the outcome variable is made after the addition of the predictor variable(s) to the model (Singer & Willett, 2003).

Table 1
Descriptive Statistics for Each of the Five Waves of SRDT Original (Raw) Scores (N = 653)

Original Variable Name	*Min*	*Max*	*M*	*SD*	*Skew*	*Kurtosis*
grade2original	12	40	36.24	3.99	-2.50	8.47
grade3original	10	45	35.99	6.18	-1.20	1.50
grade4original	8	54	41.48	7.39	-1.36	2.09
grade5original	4	54	44.46	6.47	-2.13	6.96
grade6original	11	54	41.17	8.33	-1.00	0.74

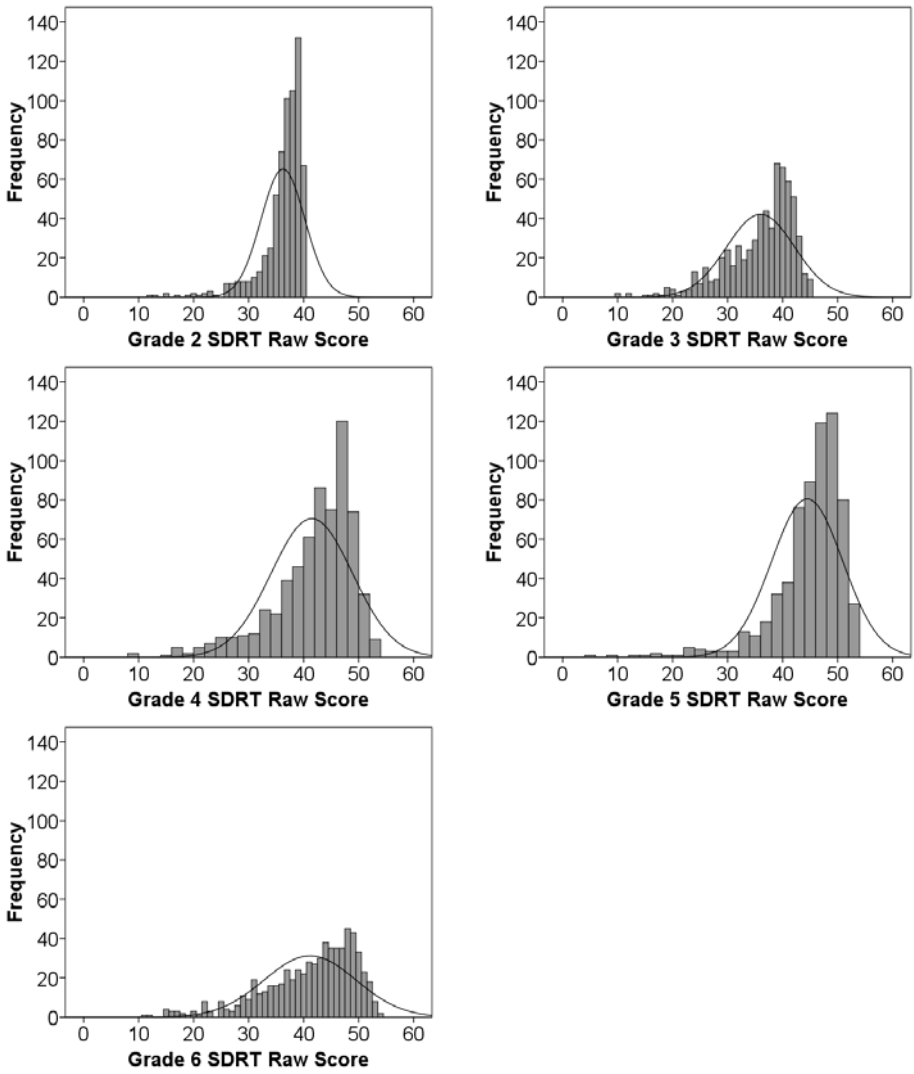

Figure 1. Histograms of the original (raw) scores across five waves: Grade 2 (top left), Grade 3 (top right), Grade 4 (middle left), Grade 5 (middle right), and Grade 6 (bottom left). Note that each of the distributions is skewed negatively.

	Example Original Variables			Resultant Rank-Transformed Variables		
	grade2 original	*grade3 original*	*grade4 original*	*grade2 rank*	*grade3 rank*	*grade4 rank*
Test-Taker *X*	30	31	29	2	1	1
Test-Taker *Y*	27	32	30	1	2.5	2
Test-Taker *Z*	36	32	35	3	2.5	3

Figure 2. An example person-level data matrix showing three waves of hypothetical original SDRT scores and their corresponding within-wave ranks.

Unconditional HLM models (sometimes called baseline or null models) generally involve computing the proportion of variance in the outcome variable that can be explained simply by the nesting of the Level 1 outcome variable (the repeated measures) within the Level 2 grouping units (the test-takers). Therefore, the Level 2 predictor variable, gender, is not included in the unconditional model.

The Level 1 model is a linear individual growth model, and represents the *within-person* (test-taker) variation. The Level 2 model expresses variation in parameters from the growth model as random effects unrelated to any test-taker level predictors (Singer, 1998), and represents the *between-person* (test-taker) variation.

Using the notation of Raudenbush and Bryk (2002), in which each level is written as a series of separate but linked equations, the relevant models and notation follow. The models and equations are expressed and formatted in a manner largely consistent with Raudenbush and Bryk (2002), Singer and Willett (2003), and Singer (1998); the *only* difference is that the dependent variable now represents test-takers' *within-wave ranks* (Y_{ti}^{R}), not the original scores. It should be noted that we present the equations in this paper primarily for the purpose of completeness, and for aiding in the interpretation of the results.

The HLM model in Equations 1 and 2 is expressed as the sum of two parts. The *fixed* part contains two fixed effects – for the intercept and for the effect of *wave* (time). The *random* part contains three random effects – for the intercept, the *wave* slope, and the within-test-taker residual (e_{ti}) (Singer, 1998).

Unconditional Model Level 1 (Within-Person)

$$Y_{ti}^{R} = \pi_{0i} + \pi_{1i}(wave)_{ti} + e_{ti,}$$

where:

- Y_{ti}^{R} = test-taker *i*'s *rank-based* literacy score on wave *t;* the superscript *R* denotes the use of ranks in the place of original scores;
- *wave* = time or measurement occasion (coded Grade 2 = 0, Grade 3 = 1, Grade 4 = 2, Grade 5 = 3, and Grade 6 = 4); (1)
- π_{0i} = test-taker *i*'s true initial status (the value of the outcome when $wave_{ti} = 0$);
- π_{1i} = test-taker *i*'s true change in relative rank during the period under study; and
- e_{ti} = the portion of test-taker *i*'s outcome that is unpredicted at wave *t* (the within-person residual).

Unconditional Model Level 2 (Between-Person)

$$\pi_{0i} = \beta_{00} + r_{0i},$$

$$\pi_{1i} = \beta_{10} + r_{1i},$$

where:

- π_{0i} = true initial status;
- π_{1i} = true change in relative rank; (2)
- β_{00} and β_{10} = Level 2 intercepts (the population average initial status and change in relative rank, respectively); and
- r_{0i} and r_{1i} = Level 2 residuals (representing those portions of initial status or rate of change that are unexplained at Level 2; in other words, they represent deviations of the individual change trajectories around their respective group average trends).

Having already fit the unconditional model in Equations 1 and 2, Equations 3 and 4 involve an HLM model which explores whether or not variation in the intercepts and slopes is related to the Level 2 predictor, gender (Singer, 1998).

Conditional Model Level 1 (Within-Person)

$$Y_{ti}^{R} = \pi_{0i} + \pi_{1i}(wave)_{ti} + e_{ti,}$$

where: (3)

- *Notation the same as in Equation 1 (Unconditional Model Level 1)*

Conditional Model Level 2 (Between-Person)

$$\pi_{0i} = \beta_{00} + \beta_{01}(gender)_{i} + r_{0i},$$

$$\pi_{1i} = \beta_{10} + \beta_{11}(gender)_{i} + r_{1i},$$

where:

- *gender* = Level 2 predictor of both initial status and change (coded male = 0 and female = 1); and (4)
- β_{01} and β_{11} = Level 2 slopes (representing the effect of gender on the change trajectories, and which provide differences in initial status and changes in relative rank, respectively).
- *All other notation the same as in Equation 2 (Unconditional Model Level 2)*

Explanation of the Statistical Output

Unconditional model. As Table 2 shows, the parameter value 327.00 represents the estimate of the average intercept across test-takers (the average value of the dependent variable when wave = 0). Therefore, the average person began with a rank of 327. The fact that this estimate is statistically significant (p = .00) simply means that this average intercept is significantly different from zero. This finding is not particularly useful, and is to be expected given that as various test-takers improve their rank over time, other test-takers worsen. Therefore, change in relative rank is a zero sum game within a closed set.[3]

Table 2 also shows that the parameter estimate for wave (0.01) is statistically non-significant ($p = 1.0$) revealing that, on average, test-takers' rank trajectories were flat. In other words, the average test-taker's rank relative to other test-takers did not change across the five waves (i.e., from Grade 2 to Grade 6).

Table 2
Unconditional Model Output: Estimates of Fixed Effects

Parameter	*Estimate*	*Standard Error*	*df*	*Test Statistic*	*p-Value*
Intercept (π_{0i})	327.00	6.52	652	50.15	.00*
Wave (π_{1i})	0.01	1.64	652	0.00	1.00

Note. * $p < .05$; Wave coded Grade 2 = 0, Grade 3 = 1, Grade 4 = 2, Grade 5 = 3, and Grade 6 = 4.

In order to postulate the appropriate individual growth model, any properties imposed on the model's composite residual (in this case, the difference between the observed and the expected value of Y^R for individual *i* on wave *t*) must match those required by the data. In specifying a model's stochastic component, one should allow for heteroscedasticity and autocorrelations among the composite residuals (Singer & Willett, 2003).

The type of covariance structure specified here for the composite residual was the *unstructured* error covariance matrix, in which each element of the hypothesised error covariance structure takes on the value demanded by the data. Because of its parsimonious nature, an unstructured error covariance matrix is viewed as being desirable for exploratory analyses (Singer & Willett, 2003) and, therefore, deemed appropriate for this paper's illustration of the *NPAR-HLM*.

Recall from Table 2 that the main effect of wave was statistically non-significant – meaning that the *average* test-taker's rank trajectory was flat. As Table 3 details, even though the average test-taker's rank relative to other test-takers did not change across the five waves (i.e., from Grade 2 to Grade 6), the variance in the intercepts (20002.27) and the slopes (471.5) are both statistically significant ($p = .01$), revealing *individual differences* in both the test-takers' relative starting rank and their change in relative rank, respectively.

Table 3
Unconditional Model Output: Estimates of Variance Parameters

Parameter	*Estimate*	*Standard Error*	*p-Value*
Variance of intercept (var(π_{0i})	20002.27	1557.8	0.01*
Variance of slopes (var(π_{1i})	471.5	106.2	0.01*

Note. * $p < .05$.

It is this variance in individual test-takers' intercepts and slopes (not the average slopes) that is particularly compelling, because this variance justifies the addition of predictor variables in Level 2 of the conditional model (in this case, gender). Recall that, in any HLM growth model, it is surmised that a given Level 2 predictor is added for the purpose of explaining differences in individual test-takers' Level 1 intercepts and slopes (Singer & Willett, 2003). Table 3, in essence, reveals that re-running the analysis with the Level 2 addition of gender is a statistically defensible next step.

Conditional model. Table 4 illustrates the output from the conditional HLM analysis. Conditional HLM models generally involve computing the proportion of variance in the outcome variable that can be explained not only by the nesting of the Level 1 scores within the Level 2 grouping units, but also by the inclusion of the predictor variable (in this case, gender) in the analysis.

The parameter estimate for *gender*, 48.87 is statistically significant (p = .00), revealing a relationship between initial status and gender. This finding suggests that female test-takers, on average, begin with a rank 48.87 higher than that of males (males coded 0, females coded 1).

A regrettable limitation of the *NPAR-HLM*, and of non-parametric solutions in general, is the difficulty it presents in terms of interpreting interactions (Sawilowsky, 1990). It is for this reason that the *wave x gender* interaction term, although presented in Table 4 for completion, is not explained here. This issue is discussed in more detail in a later section.

Strengths of the NPAR-HLM

The *non-parametric HLM* – offered as a means of handling the problem of analysing change and growth with time-variable measures collected over three or more waves – has several strengths:

Ease of use. The first strength relates to the ease of implementation of the *NPAR-HLM*. As Conover and Iman (1981) observe, it is often more convenient to use ranks

Table 4
Conditional Model Output: Estimates of Fixed Effects

Parameter	*Estimate*	*Standard Error*	*df*	*Test Statistic*	*p-Value*
Intercept (π_{0i})	301.85	9.26	651	32.57	.00*
Wave (π_{1i})	-2.46	2.35	651	-1.04	.29
Gender	48.87	12.91	651	3.78	.00*
Wave * Gender	4.79	3.28	651	1.46	.14

Note. * $p < .05$. Wave coded Grade 2 = 0, Grade 3 = 1, Grade 4 = 2, Grade 5 = 3, and Grade 6 = 4. Gender coded male = 0 and female = 1.

in a parametric statistical program than it is to write a program for a non-parametric analysis. Furthermore, all of the steps required for the implementation of the *NPAR-HLM* (i.e., rank transforming data within waves, restructuring the data matrix, conducting hierarchical/mixed-effect analyses, etc.) can be easily performed using commonly-used statistical software packages, such as SPSS and SAS.[4] Given the widespread use of such software packages in educational research settings, the *NPAR-HLM* is particularly appealing from a practical standpoint.

Bridges the parametric/non-parametric gap. Second, by rank transforming the data pre-analysis, one is able to bridge the gap between parametric and non-parametric statistical methods, thereby providing "a vehicle for presenting both the parametric and nonparametric methods in a unified manner" (Conover & Iman, 1981, p. 128). Unfortunately, introductory statistics courses and textbooks very often treat the two methods as if they are completely distinct from one another when, in fact, there can be great strength in marrying the two (Conover, 1999; Conover & Iman, 1981).

Makes use of the ordinal nature of data. Third, the *NPAR-HLM* makes use of the inherent ordinal nature of continuous-scored data: A test-taker with a *low* original score relative to other test-takers in his wave will also yield a *low* relative rank. Similarly, a test-taker with a *high* test-score will also yield a *high* rank. As a result, within-wave order among the test-takers is preserved. By ranking test-takers' scores within-wave pre-analysis, it is possible to put the longitudinal test scores onto a common metric, thereby providing a standard against which test-takers' scores can be measured and compared.

Requires no common/linkable items. Unlike many other test linking methods and strategies, the *NPAR-HLM* can be implemented not only in situations in which one's study involves time-variable measures that can be linked, but also situations in which the time-variable measures share *no linkable items whatsoever*. Hence, unlike vertical scaling, equating, and their linking counterparts, the *NPAR-HLM* provides a means by which researchers can study change and growth – whether or not the measures contain linkable items. It is anticipated that this particular feature of the *NPAR-HLM* will likely prove appealing to researchers studying constructs thought to change developmentally (e.g., academic achievement).

Requires no norming group. Recall from an earlier section that a norming group (normative sample) is a large sample, ideally representative of a well-defined population, whose test scores provide a set of standards against which the scores of one's sample can be compared (Gall, Borg, & Gall, 1996). Due to time and financial constraints, it is not always possible to compare the scores of one's sample to those of an external norming sample. As such, a fifth strength of the *NPAR-HLM* is that it can be conducted using simply the scores of the immediate sample of test-takers.

Limitations of the NPAR-HLM

As with any methodological tool, the *NPAR-HLM* has various limitations:

Problems interpreting interactions. As described in an earlier section, a regrettable limitation of the *NPAR-HLM*, and of non-parametric methods in general, is the difficulty it presents in terms of interpreting interactions. Sawilowsky (1990) describes the problems one can encounter when substituting ranks in two- or higher-way designs. For example, Sawilowsky, Blair, and Higgins (1989) found that the null hypothesis for interaction can actually change from false to true when one substitutes ranks in a finite population which can, in turn, have a negative impact on power. It is for such reasons that the *NPAR-HLM* is best suited for exploring main effects, rather than interactions. For the specific situations in which ranks may be used to detect interactions, please refer to Thompson (1991).

Within-wave ranks are bounded. As described earlier, *rank transforming* refers to the process of converting a test-taker's original score to rank relative to other test-takers – suggesting a one-to-one function *f* from the set $\{X_1, X_2, \ldots, X_N\}$ (the sample values) to the set $\{1, 2, \ldots, N\}$(the first *N* positive integers) (Zimmerman & Zumbo, 1993). The values assigned by the function to each sample value in its domain are the number of sample values having lesser or equal magnitude. Consequently, the ranks are bounded from above by *N*.

Imagine that, on a standardised test of intelligence, Test-Taker *W* earns a score 100, Test-Taker *X* earns a score of 101, Test-Taker *Y* earns a score of 102, and Test-Taker *Z* earns a score of 167. Test-Taker *Z*'s score, relative to the other test-takers, is exceptional. Despite her exceptional performance on the measure, however, Test-

Taker *Z*'s test score is masked by the application of ranks: Test-taker $W = 1$, Test-taker $X = 2$, Test-taker $Y = 3$, and Test-taker $Z = 4$. As a result, one limitation of the Conover solution is that there may be problems associated with the inherent restriction of range it places on data. Differences between any two ranks range between 1 and $N - 1$, whereas the differences between original sample values range between 0 and infinity (Zimmerman & Zumbo, 1993).

It should be noted that this limitation is also a strength because as Zimmerman and Zumbo (1993) note, as a result, "any outliers among the original sample values are not represented by deviant values in the rank" (p. 487) making the *NPAR-HLM* less sensitive to outlying data points within a wave of data, than the more typically-used parametric version of HLM growth models.

Difficulties associated with handling missing data. Recall from an earlier section that only those test-takers for whom data were available at each and every wave were retained in the analyses. As most educational researchers will note, no discussion about change and growth is complete without a complementary discussion about one unavoidable problem: missing data. In longitudinal designs, particularly those that span months or years, it is extremely common to face problems associated with participant dropout, attrition, as well as participants who join, or return to the study, in later waves.

The complexity (even messiness!) of many longitudinally-collected data sets can have serious implications for growth analyses. Singer and Willett (2003) remind readers that, when fitting a growth model:

> You implicitly assume that each person's observed records are a random sample of data from his or her true growth trajectory. If your design is sound, and everyone is assessed on every planned occasion, your observed data will meet this assumption. If one or more individuals are not assessed on one or more occasions, your observed data may not meet this assumption. In this case, your parameter estimates may be biased and your generalizations incorrect (p. 157).

One possible strategy for circumventing, or at least mitigating the effect of, missing data is to impute the missing original or standardised scores *prior to* rank-transforming the data within-wave pre-analysis. Because detailed discussion of various imputation methods are beyond the scope of this paper, and because missing data discussion is largely case-dependent, readers are referred to Schumacker and Lomax (2004) for discussion about handling missing data.

Makes use of the ordinal nature of data. Recall that the fact that the *NPAR-HLM* makes use of the ordinal nature of continuous-scored data was previously identified as one of the solution's strengths. Unfortunately, precisely what the *NPAR-HLM* wins by, it also loses by. Because of the rank transformation of the original or standardised scores:

> ... differences between raw scores are not necessarily preserved by the corresponding ranks. For example, a difference between the raw scores corresponding to the 15th and the 16th ranks is not necessarily the same as the difference between the raw scores corresponding to the 61st and 62nd ranks in a collection of 500 test scores (Zimmerman & Zumbo, 2005, p. 618).

Conclusions

There are two primary reasons why investigating the problem of analysing change and growth with time-variable measures was undertaken in this paper. First, as Willett et al. (1998) and von Davier et al. (2004) describe, the rules about which tests are permissible for repeated measures designs are precise and strict. Given these conditions, it was necessary to investigate if and how repeated measures designs are possible – speaking both psychometrically and practically – when the measures themselves must change across waves. Second, given the substantial growth in longitudinal large-scale achievement testing (Braun, 1988), it was – and is – necessary to find viable and coherent solutions to the problem so that researchers, educational organisations, policy makers, and testing companies can make the most accurate inferences possible about their test scores.

Recall from an earlier section that it is not possible to explore growth in performance outcomes when one is dealing with measures that cannot be linked or equated (as is the case with many time-variable measures), because the various measures' scores have not been placed onto any sort of common metric and, as a result, there is no way of interpreting the original scores in any meaningful way. To this end, readers were introduced to a novel solution for handling the problem of analysing change and growth with time-variable measures, particularly those that cannot be equated or linked.

It should, however, be stressed that the *NPAR-HLM* is by no means a universal panacea. As Linn (1993) notes, considering any one individual method as the ultimate solution to the problem of linking test scores is fundamentally unsound because:

> The sense in which the scores for individual test-takers can be said to be comparable to each other or to a fixed standard depends fundamentally on the similarity of the assessment tasks, the conditions of administration, and their cognitive demands. The strongest inferences that assume the interchangeability of scores demand high degrees of similarity. Scores can be made comparable in a particular sense for assessments that are less similar. Procedures that make scores comparable in one sense (e.g., the most likely score for a student on a second assessment) will not simultaneously make the scores comparable in another sense (e.g., the proportion of test-takers that exceed a fixed standard). Weaker forms of linkage are likely to be context, group, and time dependent,

which suggests the need for continued monitoring of the comparability of scores (p. 100).

Because of the *case-specific* nature of the problem of analysing change and growth with time-variable measures (that can or cannot be linked), researchers are beseeched to prioritise the making of careful and trained judgements about their proposed measures – right at the outset of the study. The later one waits to make such judgements, the less accurate the inferences one makes from the measures' scores. In conclusion, readers are advised to be mindful of the words of Kolen and Brennan (2004): "The more accurate the information, the better the decision" (p. 2).

Endnotes

1. One can also assign ranks so that the test-taker with the highest score receives a rank of 1. It is, however, easier to think of test-takers receiving the highest score also receiving the highest rank value.

2. Lloyd (2006) discusses each of these assumptions in greater detail.

3. We thank an earlier reviewer for this observation.

4. Hierarchical analyses can also be performed using such statistical packages as HLM and MLwiN; however, these packages' current versions are not able to convert original scores to ranks, so the rank transformation must be done in another statistical package pre-analysis.

References

Beasley, T. M., & Zumbo, B. D. (2003). Comparison of aligned Friedman rank and parametric methods for testing interactions in split-plot designs. *Computational Statistics and Data Analysis, 42,* 569-593.

Braun, H. I. (1988). A new approach to avoiding problems of scale in interpreting trends in mental measurement data. *Journal of Educational Measurement, 25*(3), 171-191.

Conover, W. J. (1999). *Practical nonparametric statistics* (3rd ed.). New York: John Wiley & Sons.

Conover, W. J., & Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician, 35,* 124-129.

Gall, M. D., Borg, W. R., & Gall, J. P. (1996). *Educational research: An introduction* (6th ed.). White Plains, NY: Longman.

Karlsen, B., & Gardner, E. F. (1978-1996). *Stanford Diagnostic Reading Test, Fourth Edition.* San Antonio, TX: Harcourt Brace Educational Measurement.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer-Verlag.

Lesaux, N. K., & Siegel, L. S. (2003). The development of reading in children who speak English as a Second Language (ESL). *Developmental Psychology, 39*(6), 1005-1019.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education, 6*(1), 83-102.

Lloyd, J. E. V. (2006). *On modelling change and growth when the measures themselves change across waves: Methodological and measurement issues and a novel non-parametric solution.* Unpublished doctoral dissertation, University of British Columbia.

Lloyd, J. E. V., & Zumbo, B. D. (2007). The non-parametric difference score: A workable solution for analysing two-wave change when the measures themselves change across waves. *Journal of Modern Applied Statistical Methods, 6*, 413-420.

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects.* Princeton, NJ: Educational Testing Service.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA: Sage Publications.

Sawilowsky, S. S. (1990). Nonparametric tests of interaction in experimental design. *Review of Educational Research, 60*(1), 91-126.

Sawilowsky, S. S., Blair, R. C., & Higgins, J. J. (1989). An investigation of the Type I error and power properties of the rank transform procedure in Factorial ANOVA. *Journal of Educational Statistics, 14*(3), 255-267.

Schumacker, R. E, & Lomax, R. G. (2004). *A beginner's guide to structural equation modeling* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics, 24,* 323-355.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* New York: Oxford Press.

Thompson, G. L. (1991). A note on the rank transform for interactions. *Biometrika, 78*(3), 697-701.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The Kernel method of equating.* New York: Springer.

Willett, J. B., Singer, J. D., & Martin, N. C. (1998). The design and analysis of longitudinal studies of development and psychopathology in context: Statistical models and methodological recommendations. *Development and Psychopathology, 10,* 395-426.

Zimmerman, D. W., & Zumbo, B. D. (1993). Relative power of parametric and nonparametric statistical methods. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences, Volume 1: Methodological issues* (pp. 481-517). Hillsdale, NJ: Lawrence Erlbaum.

Zimmerman, D. W., & Zumbo, B. D. (2005). Can percentiles replace raw scores in statistical analysis of test data? *Educational and Psychological Measurement, 65,* 616-638.

Zumbo, B. D. (1999). The simple difference score as an inherently poor measure of change: Some reality, much mythology. In Bruce Thompson (Ed.). *Advances in social science methodology, Volume 5,* (pp. 269-304). Greenwich, CT: JAI Press.